DOCUMENT RESUME

ED 029 364 EA 002 144

By-Skager, Rodney W.

Student Entry Skills and the Evaluation of Instructional Programs: A Case Study.

Pub Date Feb 69

Note-17p.: Paper presented at the Annual Meeting of the Amer. Educ. Res. Assn. (Los Angeles, Calif., Feb. 5-8, 1969)

Available from-Center for the Study of Evaluation of Instructional Programs, Univ. of California at Los Angeles, Los Angeles, Calif. 90024 (no charge)

EDRS Price MF-$0.25 HC Not Available from EDRS.

Descriptors-Academic Achievement, *Course Objectives, Curriculum Development, Effective Teaching, Evaluation Methods, Grade 7, *Junior High School Students, Low Achievers, Mathematics, *Pretesting, *Program Evaluation, *Skill Development, Test Validity

Identifiers-Diagnostic Tests (Education)

To investigate the assertion that there is a tendency for teachers to emphasize skills already mastered by their students, data were collected on 488 seventh grade mathematics students in three junior high schools. Of these, 285 were assigned to experimental classes taking the curriculum under development and 203 were assigned to comparison classes providing the regular mathematics curriculum. The students were average in intelligence, but were at least 1 year behind in mathematics achievement. The main testing instrument was the Diagnostic Test with which the students were both pre- and posttested. Relevancy ratings were collected to check on the fairness or appropriateness of the test items. The results showed that (1) teachers selected instructional objectives that reflected skills already available to their students, and (2) experimental teachers tended to gear instruction to skills already achieved by students at entry into the program. Two implications are drawn: (1) With regard to instructional practice, teachers need to be informed about the entry skills of their students as related to the objectives of a course of instruction; and (2) data on entry skills could have been most useful to teachers developing the program had they been made available early in the research. (HW)

ED029364

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

STUDENT ENTRY SKILLS AND THE EVALUATION* OF INSTRUCTIONAL PROGRAMS: A CASE STUDY

Rodney W. Skager

University of California at Los Angeles

and

Center for the Study of Evaluation

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL BY MICROFICHE ONLY HAS BEEN GRANTED BY Rodney W. Skager TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U.S. OFFICE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

It is axiomatic among curriculum experts that teachers often fail to acquaint themselves with the patterns of skills students bring initially to their classrooms. When new instructional programs are being developed for later use by large numbers of teachers such failure to monitor entry skills can result in a grossly inadequate match between the learning needs of students and the content of instruction. The present research goes further, however, and suggests that under certain circumstances there may even be a tendency for teachers to emphasize skills already mastered by their students. This paper first presents empirical evidence for such an assertion and then attempts to deduce some of the reasons why teachers under conditions similar to those encountered in this research might direct instruction at the improvement of skills already attained by a majority of their students.

Students and Instructional Programs -- Findings reported in this research are based on data collected as a part of the evaluation of a curriculum development program in seventh grade mathematics. For the analyses reported here data were available

* This is a draft being circulated for review. It is not to be quoted without permission of the author.
(Paper presented at annual meeting of AERA, Los Angeles, Feb. 5-8, 1969).

Copies available without charge from the Center for the Study of Evaluation of Instructional Programs, University of California at Los Angeles, Los Angeles, California 90024

ED029364

on 488 students in three junior high schools. Of these, 285 were assigned to experimental classes taking the curriculum under development and 203 were assigned to comparison classes providing the regular mathematics curriculum for each school. Within each school experimental and comparison groups were not in every case equivalent at the beginning of instruction, but this fact has no bearing on the issues discussed here. The three schools were located in a metropolitan context, in two cases of an "inner" city type. Students were varied in racial and ethnic characteristics. As identified below, approximately 95% of the students in School 1 were Mexican-Americans, with an approximately equal percentage of the students in School 2 being Negro. School 3 was a more mixed ethnic character, with somewhat over 60% of the students Caucasian, about 30% Mexican-American, and a small percentage of Negro and Oriental students. The two sexes were approximately equally represented in all groups.

The majority of the students in the three schools had taken the California Test of Mental Maturity (1957 Short Form) at the end of the fifth grade. Mean total I.Q. scores at each school for experimental and comparison groups were respectively: School 1, 94 and 90; School 2, 90 for both groups; and School 3, 95 and 100. All students assigned by the schools to either experimental or comparison classes were at least one year behind in mathematics achievement for their own school. In general, then, students participating in this research can be described as in the main members of urban minority groups who have shown unsatisfactory achievement in mathematics in comparison with their peers. Their

academic performance probably cannot be accounted for by low scholastic aptitude, as means for all groups are within the normal range.

Teachers in the experimental program were volunteers selected for their high professional qualifications. With one exception, all experimental teachers were from other schools on temporary assignment to the program. Teachers of comparison classes were on the regular staff of the participating schools. Thirteen experimental and thirteen comparison teachers contributed data to the present research.

The experimental teachers spent only half of their time in the classroom, using the rest of the day for program development activities. In general, the experimental programs at the three schools utilized programmed materials, games, and modern media in the presentation of content that was to some extent oriented to the "modern" math. While such elements were by no means excluded from the comparison mathematics classes, the latter were less richly supplied with materials and in most cases placed more emphasis on the development of computational skills. The experimental programs developed at the three schools were by no means identical and all data presented below are broken down by group within school.

Testing -- Students in both experimental and comparison programs were administered a variety of tests and other measures, though only the Diagnostic Test is of interest here. This instrument was constructed for the evaluation research of which this study is a part and was made up of items judged to be pertinent

to the instructional goals of the experimental program. The possibly different instructional goals of the comparison program were not taken into consideration except for the inclusion of a subset of computational problems in addition, subtraction, multiplication, and division. The rest of the items provided a more heterogeneous array of combinations of content and processes than would be found in the typical standardized achievement test, since one of the intentions of the research was to compare the two instructional programs on a variety of individual items or subgroups of items. Two forms of the test were developed by randomly assigning the members of a pair items of each type to Form A or Form B of the test. Students took the same form of the Diagnostic Test at the beginning and end of the school year, with forms randomly assigned to classes within each program at each school. The proportion of students passing each item on each form was calculated for pre- and posttest data for each combination of school, form of the test, and instructional program (e.g., School 1, Form A, experimental program).

Ratings of Relevancy -- After the posttesting all teachers in the two programs were given copies of the two forms of the Diagnostic Test, which they had not heretofore seen, and asked to judge the relevancy of each item to instruction in their classes during the school year. Specifically, the 13 cooperating teachers in each group were instructed to,

> "Make a judgement on the extent to which instruction in your mathematics classes this year would facilitate students' ability to answer each item correctly."

The teachers were instructed to use the following 5 point rating scale:

1 Definitely would not facilitate ability to answer

2 Probably would not facilitate

3 Uncertain

4 Probably would facilitate

5 Definitely would facilitate ability to answer

Mean ratings for teachers in each combination of test form, school, and instructional program were calculated for each item.

Purpose -- Initially the relevancy ratings were collected to provide a check on the fairness or appropriateness of the items included in the Diagnostic Test. It was hoped that all or nearly all items would be judged to be closely related to the content of instruction, especially in the experimental group. It was also of interest to determine whether the comparison teachers would see the items as less relevant than did the experimental teachers, as might be expected in view of the criteria used in selecting the items.

In examining the data, however, the author chanced to see an initial item difficulty index for one of the subgroups in juxtaposition with the mean rating of the item by the teachers of those particular students. All of the students in the group had passed the item at pretest, yet all of the teachers of those students had rated the item 5, implying a definite relevancy to instructional content! This startling observation led the author to look into the overall relationship between initial item difficulty and teacher ratings of item relevancy. For this purpose,

correlations between proportion of students passing each item at pretest and mean rating of item relevancy were calculated for each subgroup of students. Ideally, such correlations should be negative, indicating that teachers place greater emphasis on those skills in which students are initially weak. Correlations of approximately zero would suggest the lack of any systematic relationship between entry skills and instructional content, seemingly an undesirable situation. Positive correlations, of course, would be even less desirable, since such findings would suggest that instructional content is oriented to student strengths rather than weaknesses.

Results -- As indicated above, the data collected in this research were analyzed for the purpose of determining the nature of the relationship between the entry skills of students and the instructional objectives of their teachers. Before presenting evidence relating to this primary issue, two preliminary questions need to be dealt with by way of anticipating possible alternative interpretations of the results.

(1) Did the teachers in both groups judge items on the Diagnostic Test to be in general relevant to their instructional objectives and were there differences in the ratings of experimental and comparison teachers? Mean ratings averaged over the 52 items in each form of the pretest are reported in Table 1. These ratings also identified according to school and experimental vs. comparison teachers, show that as a whole the test was judged relevant to instructional goals as perceived by the teachers themselves. All but one of the means are above 3.0, the point of uncertainty.

Table 1

Mean Ratings of Relevancy of Diagnostic Test Items*

	Form A		Form B	
	Exp.	Comp.	Exp.	Comp.
School 1	3.71	4.28	3.81	4.23
	(5)	(3)	(5)	(3)
School 2	4.0	3.94	3.98	3.82
	(4)	(8)	(4)	(8)
School 3	3.61	3.59	3.36	2.95
	(4)	(2)	(4)	(2)

* Numerals in parentheses refer to number of teachers contributing to each mean rating.

Perusal of the data on individual items for each of the twelve subgroups shown above did show variability for all groups in ratings across items, but also revealed many items with mean ratings at or close to 5.0 for a given group.

It was anticipated that the experimental teachers would judge the test to be more relevant, since their planned instructional goals were the major consideration in the selection of test items. This expectation was certainly not confirmed, as the mean ratings in Table 1 reveal no pattern of differences between ratings by experimental and comparison teachers. In most cases the means are very close for the two groups within each school, and the highest mean in the table was generated by

comparison teachers at School 1.

(2) To what extent did experimental and comparison teachers agree on the relevancy of individual items of the test? While overall ratings of the relevancy of items were remarkably similar for the experimental and comparison teachers, it does not necessarily follow that teachers in the two groups saw the same items as relevant. Indeed, if all teachers gave similar relevancy ratings on each item, any differences between the two groups in the correlations of initial item difficulty with relevancy ratings could only be explained in terms of systematic differences between experimental and comparison students in the skills available at entry. Since there is no reason to believe that the process of assigning students to groups would result in systematic differences in patterns of entry skills, this explanation of the results would lead nowhere.

To answer the above question, mean ratings by experimental and comparison teachers on each items were correlated over the 52 items within school and by test form. These correlations are reported in Table 2. Inspection of these correlations reveals

Table 2

Correlations Between Mean Relevancy Ratings on Individual Test Items for Experimental and Comparison Teachers

	School 1	School 2	School 3
Form A	.35	.62	.20
Form B	.12	.64	.18

that only in the case of School 2 is there a relatively high relationship between the relevancy ratings of experimental and comparison teachers. In the case of the other two schools, the correlations, while positive, are quite weak. A possible interpretation of this finding may lie in the report by members of the evaluation staff assigned to the schools as periodic observers that only at School 2 was either formal or informal discussions between experimental and comparison teachers about instructional objectives known to have occurred. Although the Diagnostic Test appears to be based on reasonably appropriate overall content for both experimental and comparison classes, it appears that different subgroups of items were seen as relevant by experimental and comparison teachers in two of the schools, with moderate positive relationships at the third school. With the above in mind the primary question posed in this paper can be dealt with.

(3) What relationship pertained between the entry skills of students as reflected in initial item difficulty and ratings by teachers of the instructional relevancy of those items? Correlations between initial item difficulty across the 52 items on each form are reported in Table 3 for each of the twelve subgroups.

Table 3

Correlations Between Proportion Answering Each Item of Diagnostic Test Correctly at Pretest and Teacher Ratings of Item Relevancy

	Form A		Form B	
	Exp.	Comp.	Exp.	Comp.
School 1	.25	.14	.62	.30
School 2	.44	.27	.58	.32
School 3	.53	.10	.62	.25

Two conclusions are immediately apparent in this table. First and most important, all of the correlations are positive, utterly confounding the seemingly reasonable expectation that the signs of the coefficients would be negative. Table 3 reveals very clearly that the larger the proportion of students able to answer each item correctly at the beginning of the year, the more likely were teachers to rate that item as highly relevant to their instruction. In short, by their own reports, the teachers appeared to have selected instructional objectives that to a considerable extent reflected skills already available to their students.

The second conclusion is also surprising. Without exception, correlations are higher for experimental groups than for corresponding comparison groups for each combination of school and test form. For Form A the average $\underline{r}$ for experimental groups across schools is .42 as compared to .17 for comparison classes. For Form B the average experimental group correlation is .61 as against .29 for comparison students. There thus appears to have been a greater tendency among experimental teachers to gear instruction to skills already achieved by students at entry into the program.

It may also be noted, incidentally, that the correlations in Table 3 are invariably higher for Form B of the test. This trend can probably be ignored, as the author neglected to control for order effects when the ratings were collected, with the result that the items on Form A were always rated first. The correlations for Form B are perhaps more accurate estimates in the sense that the judges were more practiced.

Discussion -- How are these results, so inconsistent with what seems to be a reasonable expectation, to be explained and what are their implications for the development and evaluation of instructional programs? We are, of course, dealing here with correlational research designed to identify relationships existing in the data rather than to explain the origin of relationships as would be the case under the conditions of an experiment. For this reason, and because of the possible importance of these findings with regard to educational practice, several alternative explanations need to be considered.

A first explanation deserving of consideration holds that the results reported above on the relationship between entry skills and relevancy ratings are spurious in the sense that the teachers could actually have emphasized different content in the classroom than was indicated by their ratings. That is, perhaps the ratings did not reflect what the teachers actually did, but rather the opposite, or at least something quite different. Admittedly, the motivation for such behavior is difficult to construe, but reasoning along the following lines does not appear unduly contrived. We can safely assume that teachers are sensitive about evaluations others make of their performance as reflected in the achievement of their students. Moreover, teachers have sufficient opportunity during the year to become aware of the patterns of subject matter skills available to their students. Given this combination of desire to "look good" and knowledge of what students can and cannot do, it would be easy to claim credit via the relevancy ratings for teaching students what they already

knew in the first place.

While such an uncharitable interpretation of the results cannot be completely discounted, it does seem improbable for at least two reasons. First, there is no positive evidence for the assertion that teachers were either consciously or unconsciously distorting the actual situation in their ratings. On the contrary, there is some evidence, partly formal, and partly informal, that the ratings were honest reflections of instructional content. In another report derived from this same research Patalino (1968) found frequent instances in which greater than average gains from pretest to posttest were accompanied by higher than average relevancy ratings for subgroups of items examined separately, suggesting that more emphasis was placed on skills rated highly relevant. The subgroups of items were not formed on the basis of an analysis of the teachers' ratings (as had originally been intended) but rather because of judged similarity of content. There are also informal instances of reports by observers of students commenting to the teachers that at least some of the material was familiar.

A second interpretation of the results assumes that the ratings do reflect accurately the content emphasized by the teachers. This assumption is at least consistent with the tentative evidence just cited. Again given that they are motivated to be judged effective in their work, this interpretation asserts that teachers find it tempting to teach available skills, knowing consciously or unconsciously that their students will then appear to be performing well, especially when they are being observed by outside evaluators. This explanation would account for the

differences between the correlations in Table 3 for experimental and comparison groups in the sense that the experimental teachers were undoubtedly under greater internal pressure to succeed, since they were "master" teachers participating in an experimental program with high visibility. Not only were behavioral scientists observing their classroom and materials, but a variety of educators as well. Except for the achievement testing, comparison teachers were in a very much more typical situation with regard to visibility.

The above is plausible, but there is an additional interesting possibility. The instruction of urban, minority children who are not achieving as well as their own peers is likely to be very hard work for most teachers. Add to this the fact that experimental teachers participated in sensitivity training sessions in which it was stressed that such children are likely to associate academic aspects of school with a sense of personal failure and inadequacy. Experimental teachers were thus in effect being urged not to give the children in the program further experiences with failure. These two conditions would also account for the fact that both experimental and comparison teachers apparently directed instruction at skills already available, (it was an easier alternative than trying to teach new content), as well as to the fact that experimental teachers did so to a greater degree in order to avoid confronting their students with further failure experiences.

As plausible as the two explanations for the conclusion that the relevancy ratings reflected instructional content accurately,

(and both could be operating at the same time) one could still argue that the results do not make sense because learning would not go on at all in the schools if instructional objectives were confined to what students already knew. In reply it can be noted that, while the above correlations are not perfect relationships and do not completely exclude the possibility of some new material being introduced into the curricula studied, as it undoubtedly was, this report does not deal with students from the affluent middle classes, but with urban minority students who are already far behind in achievement and who, if Coleman's (1966) findings apply, will fall further behind as time passes. The present results are quite consistent with this well-documented phenomenon. Thus, it seems reasonable to conclude that the frustrations encountered in teaching educationally handicapped students plus the need perceived by teachers to provide such students with experiences of success plus the teachers' own needs to perform well, especially under conditions of close observation, may well lead teachers to make the task of instruction easier by emphasizing those areas of content in which present capabilities of students are relatively more developed.

Implications -- Of the two major implications of these findings, the first relates to instructional practice and the second to the methodology of evaluation. With regard to the former, it is readily apparent that teachers do need to be informed about the entry skills of their students as related to the objectives of a course of instruction, because without such information there is undue latitude for the operation of other

irrelevant factors in decisions about curriculum. Such information on entry skills will be most useful if referred to specific, unequivocal objectives such as those described by Popham (1969). The importance of obtaining information on entry skills has, of course, been stressed by others, including Glaser (1967), in the development of individualized instructional curricula.

Secondly, it is clear in the present case that data on entry skills could have been most useful to the teachers developing the program had it been made available early in the research. This failure to meet the needs of program developers is illustrative of an all too common phenomenon in evaluation. There is a widespread tendency for researchers engaged in the evaluation of instruction to concentrate on collecting data relevant to program <u>adoption</u> at the expense of data relevant to program <u>development</u>. That is, behavioral scientists typically approach evaluation of educational practices with the analogue of the experiment firmly in mind. This leads to undue concern with answering the question, "Is the new program better than the old?", and results in a neglect of the more important task of helping program developers make certain the answer will turn out to be in the affirmative. Unlike the experimenter in a controlled laboratory situation, it is highly appropriate for the educational researcher in the role of evaluator to produce data that will lead to modifications in "treatment" variables while the research is going on. As illustrated in the present case, if the evaluator does not seek out systematic information relevant to program development, others are not likely to either. This need has already

been pointed out by Cronbach (1963), Stufflebeam (1968), as well as by others. Certainly, the present research provides strong support for the assertion that all who are involved in the development and evaluation of programs of instruction should monitor the entry skills of the target population.

References

Coleman, J.S., "Equality of educational opportunity," United States Office of Education, 1966.

Cronbach, L.J., "Course improvement through evaluation," Teachers College Record, 1963, 64, p. 672-683.

Glaser, R., "Adapting the elementary school curriculum to individual performance," Proceedings of the 1967 Invitational Conference on Testing Problems, Educational Testing Service, 1967, p. 3-36.

Patalino, Marianne, "The rationale and use of content relevant achievement tests for the evaluation of instructional programs," Unpublished Master's Thesis, Graduate School of Education, UCLA, 1968.

Popham, W.J., "The controlling effect of educational objectives on curriculum," Encyclopedia of Education, in press.

Stufflebeam, D.L., "Evaluation as enlightenment for decision-making," Evaluation Center, Ohio State University, Columbus, Ohio, 1968.